Universal Dependency Annotation for Multilingual Parsing

Ryan McDonald[†] **Joakim Nivre**[†∗] **Yvonne Quirmbach-Brundage**[‡] **Yoav Goldberg**[†⋆]
Dipanjan Das[†] **Kuzman Ganchev**[†] **Keith Hall**[†] **Slav Petrov**[†] **Hao Zhang**[†]
Oscar Täckström[†∗] **Claudia Bedini**[‡] **Núria Bertomeu Castelló**[‡] **Jungmee Lee**[‡]
Google, Inc.[†] Uppsala University[∗] Appen-Butler-Hill[‡] Bar-Ilan University[⋆]
Contact: `ryanmcd@google.com`

Abstract

We present a new collection of treebanks with homogeneous syntactic dependency annotation for six languages: German, English, Swedish, Spanish, French and Korean. To show the usefulness of such a resource, we present a case study of cross-lingual transfer parsing with more reliable evaluation than has been possible before. This 'universal' treebank is made freely available in order to facilitate research on multilingual dependency parsing.[1]

1 Introduction

In recent years, syntactic representations based on head-modifier dependency relations between words have attracted a lot of interest (Kübler et al., 2009). Research in dependency parsing – computational methods to predict such representations – has increased dramatically, due in large part to the availability of dependency treebanks in a number of languages. In particular, the CoNLL shared tasks on dependency parsing have provided over twenty data sets in a standardized format (Buchholz and Marsi, 2006; Nivre et al., 2007).

While these data sets are standardized in terms of their formal representation, they are still heterogeneous treebanks. That is to say, despite them all being dependency treebanks, which annotate each sentence with a dependency tree, they subscribe to different annotation schemes. This can include superficial differences, such as the renaming of common relations, as well as true divergences concerning the analysis of linguistic constructions. Common divergences are found in the analysis of coordination, verb groups, subordinate clauses, and multi-word expressions (Nilsson et al., 2007; Kübler et al., 2009; Zeman et al., 2012).

These data sets can be sufficient if one's goal is to build monolingual parsers and evaluate their quality without reference to other languages, as in the original CoNLL shared tasks, but there are many cases where heterogenous treebanks are less than adequate. First, a homogeneous representation is critical for multilingual language technologies that require consistent cross-lingual analysis for downstream components. Second, consistent syntactic representations are desirable in the evaluation of unsupervised (Klein and Manning, 2004) or cross-lingual syntactic parsers (Hwa et al., 2005). In the cross-lingual study of McDonald et al. (2011), where delexicalized parsing models from a number of source languages were evaluated on a set of target languages, it was observed that the best target language was frequently not the closest typologically to the source. In one stunning example, Danish was the worst source language when parsing Swedish, solely due to greatly divergent annotation schemes.

In order to overcome these difficulties, some cross-lingual studies have resorted to heuristics to homogenize treebanks (Hwa et al., 2005; Smith and Eisner, 2009; Ganchev et al., 2009), but we are only aware of a few systematic attempts to create homogenous syntactic dependency annotation in multiple languages. In terms of automatic construction, Zeman et al. (2012) attempt to harmonize a large number of dependency treebanks by mapping their annotation to a version of the Prague Dependency Treebank scheme (Hajič et al., 2001; Böhmová et al., 2003). Additionally, there has been efforts to manually or semi-manually construct resources with common syn-

[1] Downloadable at https://code.google.com/p/uni-dep-tb/.

tactic analyses across multiple languages using alternate syntactic theories as the basis for the representation (Butt et al., 2002; Helmreich et al., 2004; Hovy et al., 2006; Erjavec, 2012).

In order to facilitate research on multilingual syntactic analysis, we present a collection of data sets with uniformly analyzed sentences for six languages: German, English, French, Korean, Spanish and Swedish. This resource is freely available and we plan to extend it to include more data and languages. In the context of part-of-speech tagging, universal representations, such as that of Petrov et al. (2012), have already spurred numerous examples of improved empirical cross-lingual systems (Zhang et al., 2012; Gelling et al., 2012; Täckström et al., 2013). We aim to do the same for syntactic dependencies and present cross-lingual parsing experiments to highlight some of the benefits of cross-lingually consistent annotation. First, results largely conform to our expectations of which target languages should be useful for which source languages, unlike in the study of McDonald et al. (2011). Secondly, the evaluation scores in general are significantly higher than previous cross-lingual studies, suggesting that most of these studies underestimate true accuracy. Finally, unlike all previous cross-lingual studies, we can report full labeled accuracies and not just unlabeled structural accuracies.

2 Towards A Universal Treebank

The point of departure for our 'universal' dependency representation is the Stanford typed dependencies for English (De Marneffe et al., 2006; de Marneffe and Manning, 2008), together with the tag set of Petrov et al. (2012) as the underlying part-of-speech representation. The Stanford scheme, partly inspired by the LFG framework, has emerged as a de facto standard for dependency annotation in English and has recently been adapted to several languages representing different (and typologically diverse) language groups, such as Chinese (Sino-Tibetan) (Chang et al., 2009), Finnish (Finno-Ugric) (Haverinen et al., 2010), Persian (Indo-Iranian) (Seraji et al., 2012), and Modern Hebrew (Semitic) (Tsarfaty, 2013). Its widespread use and proven adaptability makes it a natural choice for our endeavor, even though additional modifications will be needed to capture the full variety of grammatical structures in the world's languages.

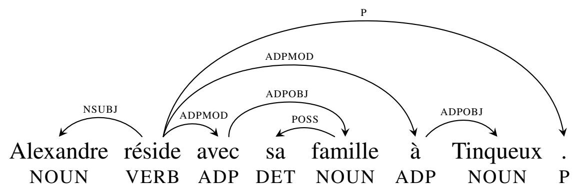

Figure 1: A sample French sentence.

We use the so-called *basic* dependencies (with punctuation included), where every dependency structure is a tree spanning all the input tokens, because this is the kind of representation that most available dependency parsers require. A sample dependency tree from the French data set is shown in Figure 1. We take two approaches to generating data. The first is traditional manual annotation, as previously used by Helmreich et al. (2004) for multilingual syntactic treebank construction. The second, used only for English and Swedish, is to automatically convert existing treebanks, as in Zeman et al. (2012).

2.1 Automatic Conversion

Since the Stanford dependencies for English is the starting point for our universal annotation scheme, we begin by describing the data sets produced by automatic conversion. For English, we used the Stanford parser (v1.6.8) (Klein and Manning, 2003) to convert the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993) to basic dependency trees, including punctuation and with the copula verb as head in copula constructions. For Swedish, we developed a set of deterministic rules for converting the Talbanken part of the Swedish Treebank (Nivre and Megyesi, 2007) to a representation as close as possible to the Stanford dependencies for English. This mainly consisted in relabeling dependency relations and, due to the fine-grained label set used in the Swedish Treebank (Teleman, 1974), this could be done with high precision. In addition, a small number of constructions required structural conversion, notably coordination, which in the Swedish Treebank is given a Prague style analysis (Nilsson et al., 2007). For both English and Swedish, we mapped the language-specific part-of-speech tags to universal tags using the mappings of Petrov et al. (2012).

2.2 Manual Annotation

For the remaining four languages, annotators were given three resources: 1) the English Stanford guidelines; 2) a set of English sentences with Stanford dependencies and universal tags (as above); and 3) a large collection of unlabeled sentences randomly drawn from newswire, weblogs and/or consumer reviews, automatically tokenized with a rule-based system. For German, French and Spanish, contractions were split, except in the case of clitics. For Korean, tokenization was more coarse and included particles within token units. Annotators could correct this automatic tokenization.

The annotators were then tasked with producing language-specific annotation guidelines with the expressed goal of keeping the label and construction set as close as possible to the original English set, only adding labels for phenomena that do not exist in English. Making fine-grained label distinctions was discouraged. Once these guidelines were fixed, annotators selected roughly an equal amount of sentences to be annotated from each domain in the unlabeled data. As the sentences were already randomly selected from a larger corpus, annotators were told to view the sentences in order and to discard a sentence only if it was 1) fragmented because of a sentence splitting error; 2) not from the language of interest; 3) incomprehensible to a native speaker; or 4) shorter than three words. The selected sentences were pre-processed using cross-lingual transfer taggers (Das and Petrov, 2011) and parsers (McDonald et al., 2011).

The annotators modified the pre-parsed trees using the TrEd[2] tool. At the beginning of the annotation process, double-blind annotation, followed by manual arbitration and consensus, was used iteratively for small batches of data until the guidelines were finalized. Most of the data was annotated using single-annotation and full review: One annotator annotating the data and another reviewing it, making changes in close collaboration with the original annotator. As a final step, all annotated data was semi-automatically checked for annotation consistency.

2.3 Harmonization

After producing the two converted and four annotated data sets, we performed a harmonization step, where the goal was to maximize consistency of annotation across languages. In particular, we wanted to eliminate cases where the same label was used for different linguistic relations in different languages and, conversely, where one and the same relation was annotated with different labels, both of which could happen accidentally because annotators were allowed to add new labels for the language they were working on. Moreover, we wanted to avoid, as far as possible, labels that were only used in one or two languages.

In order to satisfy these requirements, a number of language-specific labels were merged into more general labels. For example, in analogy with the *nn* label for (element of a) noun-noun compound, the annotators of German added *aa* for compound adjectives, and the annotators of Korean added *vv* for compound verbs. In the harmonization step, these three labels were merged into a single label *compmod* for modifier in compound.

In addition to harmonizing language-specific labels, we also renamed a small number of relations, where the name would be misleading in the universal context (although quite appropriate for English). For example, the label *prep* (for a modifier headed by a preposition) was renamed *adpmod*, to make clear the relation to other modifier labels and to allow postpositions as well as prepositions.[3] We also eliminated a few fine distinctions in the original Stanford scheme that were not annotated consistently across languages (for example, merging *complm* with *mark*, *number* with *num*, and *purpcl* with *advcl*). The final set of labels is listed with explanations in Table 1. Note that relative to the universal POS tagset of Petrov et al. (2012) our final label set is quite rich (40 versus 12). We

Label	Description	Label	Description
acomp	adjectival comp.	det	determiner
adp	adposition	dobj	direct object
adpcomp	comp. of adp.	expl	expletive
adpmod	adpl. mod.	infmod	infinitival mod.
adpobj	object of adp.	iobj	indirect object
advcl	adverbial clause mod.	mark	marker
advmod	adverbial mod.	mwe	multi-word expression
amod	adjectival mod.	neg	negation
appos	appositive	nmod	noun mod.
attr	attribute	nsubj	nominal subject
aux	auxiliary	nsubjpass	passive nominal subject
auxpass	passive auxiliary	num	numeric mod.
cc	conjunction	p	punctuation
ccomp	clausal comp.	parataxis	parataxis
compmod	compound mod.	partmod	participial mod.
conj	conjunct	poss	possessive
cop	copula	prt	verb particle
csubj	clausal subject	rcmod	relative clause mod.
csubjpass	passive clausal subject	rel	relative
dep	generic	xcomp	open clausal comp.

Table 1: Harmonized label set based on Stanford dependencies (De Marneffe et al., 2006). *adp.* = adposition, *adpl.* = adpositional, *mod.* = modifier, *comp.* = complement.

[2]http://ufal.mff.cuni.cz/tred/

[3]Consequently, *pobj* and *pcomp* were changed to *adpobj* and *adpcomp*.

Source Training Language	Target Test Language											
	Unlabeled Attachment Score (UAS)						Labeled Attachment Score (LAS)					
	Germanic			Romance			Germanic			Romance		
	DE	EN	SV	ES	FR	KO	DE	EN	SV	ES	FR	KO
DE	74.86	55.05	65.89	60.65	62.18	40.59	64.84	47.09	53.57	48.14	49.59	**27.73**
EN	58.50	83.33	**70.56**	68.07	70.14	**42.37**	48.11	78.54	**57.04**	56.86	58.20	26.65
SV	**61.25**	**61.20**	80.01	67.50	67.69	36.95	**52.19**	**49.71**	70.90	54.72	54.96	19.64
ES	55.39	58.56	66.84	78.46	**75.12**	30.25	45.52	47.87	53.09	70.29	**63.65**	16.54
FR	55.05	59.02	65.05	**72.30**	81.44	35.79	45.96	47.41	52.25	**62.56**	73.37	20.84
KO	33.04	32.20	27.62	26.91	29.35	71.22	26.36	21.81	18.12	18.63	19.52	55.85

Table 2: Cross-lingual transfer parsing results. Bolded are the best per target cross-lingual result.

	source(s)	# sentences	# tokens
DE	N, R	4,000	59,014
EN	PTB*	43,948	1,046,829
SV	STB†	6,159	96,319
ES	N, B, R	4,015	112,718
FR	N, B, R	3,978	90,000
KO	N, B	6,194	71,840

Table 3: Data set statistics. *Automatically converted WSJ section of the PTB. For the data release, will we include scripts to generate this data, not the data itself. †Automatically converted Talbanken section of the Swedish Treebank. N=News, B=Blogs, R=Consumer Reviews.

are afforded such a rich label set as the majority of the data is being hand annotated and not converted from treebanks. In the latter, one is forced to essentially work at the granularity of the coarsest treebank. If needed, our label set can easily be mapped to a coarser one for increased generality.

2.4 Final Data Sets

Table 3 presents the final data statistics. The number of sentences, tokens and tokens/sentence vary due to the source and tokenization. For example, Korean has 50% more sentences than Spanish, but ∼40k less tokens due to a more coarse-grained tokenization. In addition to the data itself, annotation guidelines and harmonization rules will be included so that the data can be regenerated.

3 Experiments

One of the motivating factors in creating such a data set was improved cross-lingual transfer evaluation. To test this, we use a cross-lingual transfer parser similar to that of McDonald et al. (2011). In particular, it is a perceptron-trained shift-reduce parser with a beam of size 8. We use the features of Zhang and Nivre (2011), except that all lexical identities are dropped from the templates during training and testing, hence inducing a 'delexicalized' model that employs only 'universal' properties from source-side treebanks, such as part-of-speech tags, labels, head-modifier distance, etc.

We ran a number of experiments, which can be seen in Table 2. For these experiments we randomly split each data set into training, development and testing sets.[4] The one exception is English, where we used the standard splits. Each row in Table 2 represents a source training language and each column a target evaluation language. We report both unlabeled attachment score (UAS) and labeled attachment score (LAS) (Buchholz and Marsi, 2006). This is likely the first reliable cross-lingual parsing evaluation. In particular, previous studies could not even report LAS due to differences in treebank annotations.

We can make several interesting observations. Most notably, for the Germanic and Romance target languages, the best source language is from the same language group. This is in stark contrast to the results of McDonald et al. (2011), who observe that this is rarely the case with the heterogenous CoNLL treebanks. Among the Germanic languages, it is interesting to note that Swedish is the best source language for both German and English, which makes sense from a typological point of view, because Swedish is intermediate between German and English in terms of word order properties. For Romance languages, the cross-lingual parser is approaching the accuracy of the supervised setting, confirming that for these languages much of the divergence is lexical and not structural, which is not true for the Germanic languages. Finally, Korean emerges as a very clear outlier (both as a source and as a target language), which again is supported by typological considerations as well as by the difference in tokenization.

With respect to evaluation, it is interesting to

[4]These splits will be included in the release of the data.

compare the absolute numbers to those reported in McDonald et al. (2011) for the languages common to both studies (DE, EN, SV and ES). In that study, UAS was in the 38–68% range, as compared to 55–75% here. For Swedish, we can even measure the difference exactly, because the test sets are the same, and we see an increase from 58.3% to 70.6%. This suggests that most cross-lingual parsing studies have underestimated accuracies.

4 Conclusion

We have released data sets for six languages with consistent dependency annotation. After the initial release of data, we will continue to annotate data in more languages as well as investigate further automatic treebank conversions. This may also lead to modifications of the annotation scheme, which should be regarded as preliminary at this point. Specifically, with more typologically and morphologically diverse languages being added to the collection, it may be advisable to consistently enforce the principle that content words take function words as dependents, which is currently violated in the analysis of adpositional and copula constructions. This will ensure a consistent analysis of functional elements that in some languages are not realized as free words. It will also allow the inclusion of language-specific functional or morphological markers (such as case markers) at the leaves of the tree, where they can easily be ignored in applications that require a uniform cross-lingual representation. Finally, we plan to make this data available on an open source repository in the hope that the community will commit new data and make corrections to existing annotations.

Acknowledgments

Many people played critical roles in the process of creating the resource. At Google, Fernando Pereira, Alfred Spector, Kannan Pashupathy, Michael Riley and Corinna Cortes supported the project and made sure it had the required resources. Jennifer Bahk and Dave Orr helped coordinate the necessary contracts. Andrea Held, Supreet Chinnan, Elizabeth Hewitt, Tu Tsao and Leigha Weinberg made the release process smooth. Michael Ringgaard, Andy Golding, Terry Koo, Alexander Rush and many others provided technical advice. Hans Uszkoreit gave us permission to use a subsample of sentences from the Tiger Treebank (Brants et al., 2002), the source of the news domain for our German data set. Annotations were additionally provided by Sulki Kim, Patrick McCrae, Laurent Alamarguy and Héctor Fernández Alcalde.

References

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank: A three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Kluwer.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL.*

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15*, pages 1–7. Association for Computational Linguistics.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with Chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, pages 51–59.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Marie-Catherine De Marneffe, Bill MacCartney, and Chris D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Tomaz Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46:131–142.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of ACL-IJCNLP*.

Douwe Gelling, Trevor Cohn, Phil Blunsom, and Joao Graça. 2012. The pascal challenge on grammar induction. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 64–80. Association for Computational Linguistics.

Jan Hajič, Barbora Vidova Hladka, Jarmila Panevová, Eva Hajičová, Petr Sgall, and Petr Pajas. 2001. Prague Dependency Treebank 1.0. LDC, 2001T10.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking finnish. In *Proceedings of The Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.

Stephen Helmreich, David Farwell, Bonnie Dorr, Nizar Habash, Lori Levin, Teruko Mitamura, Florence Reeder, Keith Miller, Eduard Hovy, Owen Rambow, and Advaith Siddharthan. 2004. Interlingual annotation of multilingual text corpora. In *Proc. of the HLT-EACL Workshop on Frontiers in Corpus Annotation*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proc. NAACL*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. of ACL*.

Dan Klein and Chris D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proceedings of ACL*.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan and Claypool.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*.

Jens Nilsson, Joakim Nivre, and Johan Hall. 2007. Generalizing tree transformations for inductive dependency parsing. In *Proc. of ACL*.

Joakim Nivre and Beáta Megyesi. 2007. Bootstrapping a Swedish treebank using cross-corpus harmonization and annotation projection. In *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of EMNLP-CoNLL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*.

Mojgan Seraji, Beáta Megyesi, and Nivre Joakim. 2012. Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7:1–10.

David A. Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of EMNLP*.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the ACL*.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of stanford dependencies. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Štepánek, Zdeněk Žabokrtskỳ, and Jan Hajic. 2012. Hamledt: To parse or not to parse. In *Proc. of LREC*.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of ACL-HLT*.

Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to map into a universal pos tagset. In *Proc. EMNLP*, pages 1368–1378. Association for Computational Linguistics.